scCancerExplorer: a comprehensive database for interactively exploring single-cell multi-omics data of human pan-cancer

Changzhi Huang[1,2#], Zekai Liu[1#], Yunlei Guo[1#], Wanchu Wang[1#], Zhen Yuan[1], Yusheng Guan[1], Deng Pan[1], Zhibin Hu[1], Linhua Sun[3]*, Zan Fu[2]*, Shuhui Bian[1,4]*

[1]State Key Laboratory of Reproductive Medicine and Offspring Health, Nanjing Medical University, Nanjing 211166, China

[2]Department of General Surgery, the First Affiliated Hospital of Nanjing Medical University, Nanjing 210029, China

[3]College of Life Sciences, Nanjing Normal University, Nanjing 210023, China

[4]Collaborative Innovation Center for Cancer Personalized Medicine, School of Public Health, Nanjing Medical University, Nanjing 211166, China

[#]Co-first author

[*]Co-correspondence

Abstract

Single-cell multi-omics data is a valuable resource to promote the understanding of cancer. However, multimodal data analysis is challenging for most cancer researchers, and there is still a lack of online databases that can bridge the gap. Here we introduced scCancerExplorer, which is a user-friendly database designed for exploring single-cell multi-omics data of human pan-cancer. We processed more than 110 datasets covering single-cell multiomics data from 51 human cancer types, and provides 5 major modules and 12 useful functions for conveniently exploring those data. (1) The "Integrated multi-omics analysis" module enables users explore a gene from multimodal perspectives, such as the gene expression levels, survival analysis, transcription factor activity, and the DNA methylation levels of promoter regions. (2) With single-cell transcriptome module, users can explore the integrated pan-cancer datasets, compare the expression profiles between primary tumor and metastases, and generate customized figures. (3) With single-cell epigenome module, users can explore promoter DNA methylation levels in both normal and cancer cells, DNA methylation state at single CpG resolution, and chromatin accessibility patterns of different cell types. (4) For single-cell genome data, users can visualize the copy number variations of cancer cells for each patient. (5) Moreover, TCGA survival analysis can be performed conveniently. Users can not only group the patients according to gene expression levels, but also select patients by gene mutation states, pathological subtypes, and cancer stages. In summary, scCancerExplorer is a powerful database for convenient data mining by simple clicking, and gaining novel insights into human pan-cancer. scCancerExplorer is freely accessible at https://bianlab.cn/scCancerExplorer.

Key words: Database; human pan-cancer; single-cell multi-omics

Introduction

Cancer is a leading cause of mortality globally(1), which is characterized by heterogeneous genetic, epigenetic, and transcriptional landscapes. Over recent years, single-cell multi-omics techniques have revolutionized cancer research, offering powerful tools to dissect tumor heterogeneity and understand the intricate molecular basis of human cancers. Moreover, large amounts of single-cell multi-omics data, encompassing the genome, transcriptome, and epigenome, have been consistently generated for human cancer studies (2–38). Importantly, these studies not only provided a comprehensive atlas of human cancers, but also offered a wealth of data resources for further exploration. However, analyzing and integrating the massive single-cell multi-omics data from separate studies posed significant challenges, particularly for wet-lab researchers lacking analytical skills and adequate computing resources. Therefore, there is an urgent need for comprehensive and user-friendly platforms to bridge the gap between abundant data resources and wet-lab researchers.

Recently, growing numbers of databases related to cancer have been developed to facilitate cancer researchers. The Cancer Genome Atlas (TCGA) provided comprehensive multi-omics data and clinical information of 33 human cancer types, providing valuable resources for exploring human cancer. Furthermore, several single-cell transcriptomic databases of human cancers have emerged and provided comprehensive gene expression atlases of both the cancers and the tumor microenvironment (TME), such as CancerSEA (39), TISCH (40), and CancerSCEM (41). However, despite these advancements, integrated multi-omics analysis stands at the forefront as a rapidly growing trend, showing its pivotal role in unravelling the complexities of cancer biology and facilitating personalized therapeutic strategies. Currently, there still remains an absence of a user-friendly and integrated web tool designed for exploring single-cell multi-omics data of human pan-cancer.

Here, we introduce scCancerExplorer, an integrated database designed to provide user-

friendly functions for exploring single-cell multi-omics data of human pan-cancer. By facilitating data mining and inspiring novel insights into pan-cancer multi-omics data, scCancerExplorer aims to empower researchers to decipher the complexities of human cancer and pave the way for cancer research.

Database content and usage

Overall design

scCancerExplorer was designed to enable convenient exploration of single-cell multi-omics data of human pan-cancer, and inspired new insights into the complex landscape of human cancers. In total, multi-omics data from more than 110 datasets covering over 4 million single cells has been collected and analyzed. We provided five major modules for users, including "Integrated multi-omics analysis", "Single-cell transcriptome", "Single-cell epigenome", "Single-cell genome" and "TCGA analysis" (Figure 1). Twelve useful functions were provided, which will be described in detail below.

"Integrated multi-omics analysis" module

This module provided a multi-omics perspective to learn an interested gene quickly, including the gene expression levels of both normal and cancer tissues using TCGA data (Figure 2A), the single-cell gene expression levels of various cell types using single-cell RNA-seq data (Figure 2B), survival analysis results using TCGA data (Figure 2C), the transcription factor (TF) activity and gene activity deduced from single-cell ATAC-seq data (Figure 2D, E), and single-cell DNA methylation levels of the promoter region in both normal and cancer cells (Figure 2F). If users want to further learn the interested gene in depth, they can use other modules for each omics below.

"Single-cell transcriptome" module

Pan-cancer analysis

Users can perform quick exploration of the gene expression pattern of pan-cancer by exploring the integrated single-cell RNA-seq data of 49 human cancer types (Figure

3A). Querying the gene expression levels of multiple interested genes in multiple cell types from multiple cancer types at the same time is allowed. Additionally, users can upload their data and integrate it with the pan-cancer data in our database for personalized comparisons (Figure 3B).

Query gene expression

In addition to the above-mentioned "Pan-cancer analysis" function, users typically need explore a certain cancer type in depth, which can be met by the "Query gene expression" function. With this function, users can easily explore the gene expression patterns of an interested gene across different cell types in certain cancer type (Figure 3C), and get other related information at the same time, including the cell type identification, sample locations (e.g. normal tissue, primary tumor, and metastasis), cancer stages, fractions of each cell cluster, sex information, cell cycle phases, and the differentially expressed genes (DEGs) of each cell type (Figure 3D). Moreover, a brief introduction and links to further information for each human gene are presented.

DIY expression plot

The "DIY expression plot" function enables users to query the expression levels of multiple selected genes in multiple selected cell types and generate customized figures to visualize (Figure 4A). Multiple figure types are provided, including dot plots, heatmaps, and violin plots, which can meet most of the needs of users. Moreover, for each cell type, users can easily compare the expression levels among different sampling locations (e.g. primary tumor v.s. metastasis) (Figure 4B), different cancer stages (e.g. stage I v.s. stage III) (Figure 4C), various cancer subtypes (e.g. poorly-differentiated adenocarcinoma v.s. well-differentiated adenocarcinoma) (Figure 4D), and between male and female by simply switching options in the "Group by" select box.

Single-cell epigenome module

Focal DNA methylation plot

With this function, users can select a narrow genomic region of interest (e.g. promoters) for visualizing the DNA methylation levels at a single CpG resolution. For example, the "Tanghulu" plot displayed the demethylated region in cancer cells compared with normal cells (Figure 5A). Each row in the plot represents an individual cell, with each circle denoting a CpG site. The methylated sites are represented by black circles, while the unmethylated sites are represented by white circles.

Promoter DNA methylation plot

Users can query the DNA methylation levels of gene promoter regions at single-cell resolution, and compare them across normal tissues, primary tumors and metastases in this module (Figure 5B).

Epigenome browser

scCancerExplorer incorporates the WashU Epigenome browser (42) to facilitate the exploration of chromatin accessibility and co-accessibility patterns. Users can conveniently access specific genomic regions or genes of interest by inputting the official gene symbol or a genomic range in the search box at the top left of the browser interface. For example, we can see five potential enhancers for *LAPMT4B* in ovarian cancer cells showed higher accessibility levels compared to other cell types (Figure 5C). The browser interface allows users to zoom in/out, navigate to new regions, and customize display options such as display mode (bar plot or heatmap), RGB colors, axis scales, and text labels to suit individual preferences.

TF motif activity

In this module, users can search for an interested transcription factor and obtain the motif activity score (motif deviation z-score) across different cell types calculated using single-cell ATAC-seq data (Figure 5D).

DIY ATAC-seq plot

Similar to the "DIY expression plot" module, this module enables users to easily query the motif activities of multiple TFs across different cell types and generate customized figures based on their requirements (Figure 5E).

"Single-cell genome" module

This module allows users to investigate the single-cell copy number alterations (CNAs) of human cancers (Figure 6A-B). For example, users can select a cancer type (e.g. stomach adenocarcinoma) and select one or multiple patients (e.g. SC02 and SC07), and then generate the CNA heatmap (Figure 6A). The cancer cells are sorted by genetic lineage if the lineage information is available.

"TCGA analysis" module

Survival analysis

This module provides survival analysis for 33 cancer types using TCGA data. Notably, users can categorize patients based on multiple information, including gene expression levels (e.g. above median v.s. below median), gene mutation states (with TP53 mutation v.s. without TP53 mutation), pathological types (all types or certain subtype), and cancer stages (all stages or certain stage). For example, users can group patients by the expression levels of *TPM2* in all COAD (Colon adenocarcinoma) patients (Figure 7A), only in patients with *APC* mutation (Figure 7B), and only in patients diagnosed as adenocarcinoma (Figure 7C), respectively.

ATAC-seq browser

With the "ATAC-seq browser" function, users can explore the open chromatin regions of 23 cancer types using TCGA ATAC-seq data. They can visualize the chromatin accessibility pattern of a gene of interest. For example, the chromatin accessibility pattern of *MISP* and *SRI*, which are highly expressed in gastric cancer (16), can be explored with this function (Figure 7D-E). Additionally, users can upload their own data for comparison.

GWAS risk SNPs

Users can quickly search the cancer-associated risk SNPs provided by GWAS catalog (https://www.ebi.ac.uk/gwas/) in this module. Moreover, users can explore the epigenetic states around the risk SNPs in the "Epigenome browser" and "Focal DNA methylation plot" functions, which can inspire novel insights.

MATERIALS AND METHODS

Data collection

The single-cell transcriptome, epigenome (DNA methylation and chromatin accessibility), and genome data (copy number variation) of 51 human cancer types were collected, curated, and analyzed. Related metadata information, such as pathological subtype, sex, sample location, and clinical stage, was collected if available. The TCGA data from 33 cancer types, including the gene expression matrix, mutation information, and clinical data were obtained from the Genomic Data Commons (GDC) Data Portal (https://portal.gdc.cancer.gov). The ATAC-seq data for 23 cancer types from TCGA project was obtained from https://gdc.cancer.gov/about-data/publications/ATACseq-AWG.

Data processing

Single-cell transcriptome data processing

In principle, the cell type identity information provided by the original studies will be preferentially used. For those datasets without provided cell type identities, first, quality control of the single-cell data was performed according to the description in the original study. Then, the single-cell gene expression matrices were processed utilizing Seurat (v.4.3.0.1) (43). Briefly, for each cell, the unique molecular identifier (UMI) count of each gene was divided by the total UMI counts of the cell and multiplied by a scale factor (default or as the original study described) before natural-

log transformed. The top highly variable genes (default or as the original study described) were selected using the "FindVariableFeatures" function. The normalized data were then centered and scaled using the "ScaleData" function. Principal component analysis (PCA) was conducted on the scaled data using the highly variable genes. Subsequently, the Uniform Manifold Approximation and Projection (UMAP) was performed and employed to visualize the cell clustering result. Cell types were annotated based on available metadata or identified using canonical marker genes. Significant differentially expressed gene (DEG) between various cell types were identified using the "FindAllMarkers" function in Seurat.

Single-cell DNA methylation data processing

For each cell, only the CpG or WCG sites with DNA methylation levels < 0.1 or > 0.9 were retained. Subsequently, the single-cell data were merged by cell group. For the CpG or WCG sites covered by multiple cells, the mean DNA methylation levels of covered cells were calculated (44). The gene promoter regions were defined as 1 kb upstream and 0.5 kb downstream of the transcription start site (TSS). For each cell, the mean methylation levels within the promoter regions were calculated to represent the methylation status of promoter regions.

Single-cell ATAC-seq data processing

To maintain maximum consistency with the original study results, only the cells present in the metadata information provided by the original study were retained. For datasets lacking metadata information, low-quality cells were removed mainly based on the criteria described in the original study. For datasets that did not provide quality control criteria in the original study, only cells with minTSS > 4 and minFrags > 1000 were used for further analyses. Downstream data processing utilized the R package ArchR (v.1.0.2) (45). In general, the "Arrow" file was created according to the quality control standard and the "addDoubletScores" function of ArchR was employed to remove the data with doublet enrichment scores > 1. The resulting "Arrow" files were used to create

the "ArchRProject". The iterative latent semantic index (LSI) dimension reduction was performed using the "addIterativeLSI" function in ArchR. Four iterations were executed with a clustering resolution ranging from 0.2 to 1, utilizing 15,000 to 25,000 variable features. Subsequently, the data was compressed into a two-dimensional UMAP embedding using the "addUMAP" function from ArchR package. The enrichment of TF motifs for peaks in each cell cluster was calculated using chromVAR (v.1.20.2). Peak calling was performed on single-cell ATAC-seq data using MACS2 (v.2.2.9.1). The "getGroupBW" function from ArchR package was utilized to export bigwig files with a tile size of 100 bp based on cell type. Peak-to-peak co-accessibility were exported using the "getCoAccessibility" function and then converted to the inter.bb format using bedTOBigBed (46).

Copy number variation data processing

The single-cell copy number variation data were provided by the authors from the original studies (16,30,31).

TCGA survival analysis

The gene expression matrix, mutation information, and clinical data for 33 cancer types were obtained from TCGA project. The "lifelines" library (v.0.27.8) in Python were utilized for the survival analysis. The expression levels were categorized into "High" and "Low" groups based on the median expression value of the selected gene. The Kaplan-Meier survival curves were generated using "KaplanMeierFitter" class, and the log-rank tests were performed. Additionally, a Cox proportional hazards model was fitted using "CoxPHFitter" class to estimate the hazard ratios. The survival curves and statistical results were visualized with matplotlib (v.3.7.5) and seaborn (v.0.13.2).

TCGA RNA-seq data analysis

For TCGA RNA-seq data, the TCGAbiolinks (v2.31.3) package were utilized to extract and clean the data. Specifically, TPM (Transcripts Per Million) matrices were extracted

using the SummarizedExperiment (v1.26.1) package. Subsequently, a log2 transformation was applied to the combined data matrix using the formula $\log_2(TPM + 1)$.

ATAC-seq browser

The ATAC-seq peak files of 23 cancer types were downloaded from the TCGA project (47).

GWAS risk SNPs

The risk SNPs associated with each cancer type were downloaded from the Genome-Wide Association Study (GWAS) catalog (48).

Discussion

In recent years, single-cell multi-omics sequencing technique has greatly advanced cancer research, offering comprehensive insights into human pan-cancers. The interplay between different molecular layers is critical for cancer progression and metastasis. However, a well-curated and user-friendly platform for exploring single-cell multi-omics data is still lacking. In this context, we present scCancerExplorer as a comprehensive single-cell multi-omics database designed for wet-lab researchers and cancer biologists.

scCancerExplorer was designed to have two important features. First, by preprocessing a large number of single-cell multi-omics data and providing rich interactive tools to exploring them, we provide an important bridge between the valuable but complex resources and cancer researchers. The 5 major modules and 12 useful functions cover analyses of both single-cell genome, epigenome (chromatin accessibility and DNA methylation) and transcriptome data, and TCGA data. Second, several customizable functions were provided to meet personalized requirements of users. For example, with the "DIY expression plot" function, users can select multiple genes and cell types and

designate their order shown in the generated figures, and use the publishing-level figures in their study. Additionally, by simply switching options in the "Group by" select box, users can easily compare the expression levels among different sampling locations (e.g. primary tumor v.s. metastasis), different cancer stages, and various cancer subtypes. For TCGA survival analysis, users can not only group the patients according to gene expression levels, but also select patients by gene mutation states, pathological subtypes, and cancer stages.

In summary, scCancerExplorer is a powerful database for convenient data mining by simple clicking, and gaining novel insights into human pan-cancer. We aim to continually develop and maintain of the database for the benefits of broad cancer biologists and wet-lab researchers.

Data availability

scCancerExplorer is freely accessible at https://bianlab.cn/scCancerExplorer.

Competing interests

The authors declare that they have no competing interests.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 82173059, No. 82172956), the Young Elite Scientists Sponsorship Program by China Association for Science and Technology (CAST) (Grant No. 2020QNRC001), and research start-up funding from Nanjing Medical University (Grant No. KY116RC20200007). All authors read and approved the final manuscript.

Reference

1. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024 May;74(3):229–63.

2. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods. 2009 May;6(5):377–82.

3. Lambo S, Trinh DL, Ries RE, Jin D, Setiadi A, Ng M, et al. A longitudinal single-cell atlas of treatment response in pediatric AML. Cancer Cell. 2023 Dec;41(12):2117-2135.e12.

4. Regner MJ, Wisniewska K, Garcia-Recio S, Thennavan A, Mendez-Giraldez R, Malladi VS, et al. A multi-omic single-cell landscape of human gynecologic malignancies. Mol Cell. 2021 Dec;81(23):4924-4941.e10.

5. Cheng S, Li Z, Gao R, Xing B, Gao Y, Yang Y, et al. A pan-cancer single-cell transcriptional atlas of tumor infiltrating myeloid cells. Cell. 2021 Feb;184(3):792-809.e23.

6. Lu Y, Yang A, Quan C, Pan Y, Zhang H, Li Y, et al. A single-cell atlas of the multicellular ecosystem of primary and metastatic hepatocellular carcinoma. Nat Commun. 2022 Aug 6;13(1):4594.

7. Paulson KG, Voillet V, McAfee MS, Hunter DS, Wagener FD, Perdicchio M, et al. Acquired cancer resistance to combination immunotherapy from transcriptional loss of class I HLA. Nat Commun. 2018 Sep 24;9(1):3868.

8. Olsen RR, Ireland AS, Kastner DW, Groves SM, Spainhower KB, Pozo K, et al. ASCL1 represses a SOX9 $^{+}$ neural crest stem-like state in small cell lung cancer. Genes Dev. 2021 Jun;35(11–12):847–69.

9. Yost KE, Satpathy AT, Wells DK, Qi Y, Wang C, Kageyama R, et al. Clonal replacement of tumor-specific T cells following PD-1 blockade. Nat Med. 2019 Aug;25(8):1251–9.

10. Yang Y, Li Y, Yu H, Ding Z, Chen L, Zeng X, et al. Comprehensive landscape of resistance mechanisms for neoadjuvant therapy in esophageal squamous cell carcinoma by single-cell transcriptomics. Signal Transduct Target Ther. 2023 Aug 11;8(1):298.

11. Zhang X, Peng L, Luo Y, Zhang S, Pu Y, Chen Y, et al. Dissecting esophageal squamous-cell carcinoma ecosystem by single-cell transcriptomic analysis. Nat Commun. 2021 Sep 6;12(1):5291.

12. Tirosh I, Izar B, Prakadan SM, Wadsworth MH, Treacy D, Trombetta JJ, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science. 2016 Apr 8;352(6282):189–96.

13. Collin J, Queen R, Zerti D, Steel DH, Bowen C, Parulekar M, et al. Dissecting the Transcriptional and Chromatin Accessibility Heterogeneity of Proliferating Cone Precursors in Human Retinoblastoma Tumors by Single Cell Sequencing—Opening Pathways to New Therapeutic Strategies? Investig Opthalmology Vis Sci. 2021 May 17;62(6):18.

14. Liu Y, Zhang Q, Xing B, Luo N, Gao R, Yu K, et al. Immune phenotypic linkage between colorectal cancer and liver metastasis. Cancer Cell. 2022 Apr;40(4):424-437.e5.

15. Fan X, Lu P, Wang H, Bian S, Wu X, Zhang Y, et al. Integrated single-cell multiomics analysis reveals novel candidate markers for prognosis in human pancreatic ductal adenocarcinoma. Cell Discov. 2022 Feb 15;8(1):13.

16. Bian S, Wang Y, Zhou Y, Wang W, Guo L, Wen L, et al. Integrative single-cell multiomics analyses dissect molecular signatures of intratumoral heterogeneities and differentiation states of human gastric cancer. Natl Sci Rev. 2023 May 10;10(6):nwad094.

17. Ji AL, Rubin AJ, Thrane K, Jiang S, Reynolds DL, Meyers RM, et al. Multimodal Analysis of Composition and Spatial Architecture in Human Squamous Cell Carcinoma. Cell. 2020 Jul;182(2):497-514.e22.

18. Jerby-Arnon L, Neftel C, Shore ME, Weisman HR, Mathewson ND, McBride MJ, et al. Opposing immune and genetic mechanisms shape oncogenic programs in synovial sarcoma. Nat Med. 2021 Feb;27(2):289–300.

19. Lambrechts D, Wauters E, Boeckx B, Aibar S, Nittner D, Burton O, et al. Phenotype molding of stromal cells in the lung tumor microenvironment. Nat Med. 2018 Aug;24(8):1277–89.

20. Laughney AM, Hu J, Campbell NR, Bakhoum SF, Setty M, Lavallée VP, et al. Regenerative lineages and immune-mediated pruning in lung cancer metastasis. Nat Med. 2020 Feb;26(2):259–69.

21. Becker WR, Nevins SA, Chen DC, Chiu R, Horning AM, Guha TK, et al. Single-cell analyses define a continuum of cell state and composition changes in the malignant transformation of polyps to colorectal cancer. Nat Genet. 2022 Jul;54(7):985–95.

22. Zhang L, Li Z, Skrzypczynska KM, Fang Q, Zhang W, O'Brien SA, et al. Single-Cell Analyses Inform Mechanisms of Myeloid-Targeted Therapies in Colon Cancer. Cell. 2020 Apr;181(2):442-459.e29.

23. Zhang Y, Narayanan SP, Mannan R, Raskind G, Wang X, Vats P, et al. Single-cell analyses of renal cell cancers reveal insights into tumor microenvironment, cell of origin, and therapy response. Proc Natl Acad Sci. 2021 Jun 15;118(24):e2103240118.

24. Zhang Y, Chen H, Mo H, Hu X, Gao R, Zhao Y, et al. Single-cell analyses reveal key immune cell subsets associated with response to PD-L1 blockade in triple-negative breast cancer. Cancer Cell. 2021 Dec;39(12):1578-1593.e8.

25. Durante MA, Rodriguez DA, Kurtenbach S, Kuznetsov JN, Sanchez MI, Decatur CL, et al. Single-cell analysis reveals new evolutionary complexity in uveal melanoma. Nat Commun. 2020 Jan 24;11(1):496.

26. Wang Y, Xie H, Chang X, Hu W, Li M, Li Y, et al. Single-Cell Dissection of the Multiomic Landscape of High-Grade Serous Ovarian Cancer. Cancer Res. 2022 Nov 2;82(21):3903–16.

27. Sun Y, Wu L, Zhong Y, Zhou K, Hou Y, Wang Z, et al. Single-cell landscape of the ecosystem in early-relapse hepatocellular carcinoma. Cell. 2021 Jan;184(2):404-421.e16.

28. LeBlanc VG, Trinh DL, Aslanpour S, Hughes M, Livingstone D, Jin D, et al. Single-cell landscapes of primary glioblastomas and matched explants and cell lines show variable retention of inter- and intratumor heterogeneity. Cancer Cell. 2022 Apr;40(4):379-392.e9.

29. Granja JM, Klemm S, McGinnis LM, Kathiria AS, Mezger A, Corces MR, et al. Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. Nat Biotechnol. 2019 Dec;37(12):1458–65.

30. Bian S, Hou Y, Zhou X, Li X, Yong J, Wang Y, et al. Single-cell multiomics sequencing and analyses of human colorectal cancer. Science. 2018 Nov 30;362(6418):1060–3.

31. Zhou Y, Bian S, Zhou X, Cui Y, Wang W, Wen L, et al. Single-Cell Multiomics Sequencing

Reveals Prevalent Genomic Alterations in Tumor Stromal Cells of Human Colorectal Cancer. Cancer Cell. 2020 Dec;38(6):818-828.e5.

32. Lee HW, Chung W, Lee HO, Jeong DE, Jo A, Lim JE, et al. Single-cell RNA sequencing reveals the tumor microenvironment and facilitates strategic choices to circumvent treatment failure in a chemorefractory bladder cancer patient. Genome Med. 2020 Dec;12(1):47.

33. Darmanis S, Sloan SA, Croote D, Mignardi M, Chernikova S, Samghababi P, et al. Single-Cell RNA-Seq Analysis of Infiltrating Neoplastic Cells at the Migrating Front of Human Glioblastoma. Cell Rep. 2017 Oct;21(5):1399–410.

34. Peng J, Sun BF, Chen CY, Zhou JY, Chen YS, Chen H, et al. Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. Cell Res. 2019 Sep;29(9):725–38.

35. Van Galen P, Hovestadt V, Wadsworth Ii MH, Hughes TK, Griffin GK, Battaglia S, et al. Single-Cell RNA-Seq Reveals AML Hierarchies Relevant to Disease Progression and Immunity. Cell. 2019 Mar;176(6):1265-1281.e24.

36. Rothzerg E, Feng W, Song D, Li H, Wei Q, Fox A, et al. Single-Cell Transcriptome Analysis Reveals Paraspeckles Expression in Osteosarcoma Tissues. Cancer Inform. 2022 Jan;21:117693512211401.

37. Puram SV, Tirosh I, Parikh AS, Patel AP, Yizhak K, Gillespie S, et al. Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. Cell. 2017 Dec;171(7):1611-1624.e24.

38. Li C, Guo L, Li S, Hua K. Single-cell transcriptomics reveals the landscape of intra-tumoral heterogeneity and transcriptional activities of ECs in CC. Mol Ther - Nucleic Acids. 2021 Jun;24:682–94.

39. Yuan H, Yan M, Zhang G, Liu W, Deng C, Liao G, et al. CancerSEA: a cancer single-cell state atlas. Nucleic Acids Res. 2019 Jan 8;47(D1):D900–8.

40. Sun D, Wang J, Han Y, Dong X, Ge J, Zheng R, et al. TISCH: a comprehensive web resource enabling interactive single-cell transcriptome visualization of tumor microenvironment. Nucleic Acids Res. 2021 Jan 8;49(D1):D1420–30.

41. Zeng J, Zhang Y, Shang Y, Mai J, Shi S, Lu M, et al. CancerSCEM: a database of single-cell expression map across various human cancers. Nucleic Acids Res. 2022 Jan 7;50(D1):D1147–55.

42. Li D, Purushotham D, Harrison JK, Hsu S, Zhuo X, Fan C, et al. WashU Epigenome Browser update 2022. Nucleic Acids Res. 2022 Jul 5;50(W1):W774–81.

43. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021 Jun;184(13):3573-3587.e29.

44. Liu Z, Yuan Z, Guo Y, Wang R, Guan Y, Chen Y, et al. SMARTdb: An Integrated Database for Exploring Single-cell Multi-omics Data of Reproductive Medicine.

45. Granja JM, Corces MR, Pierce SE, Bagdatli ST, Choudhry H, Chang HY, et al. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. Nat Genet. 2021 Mar;53(3):403–11.

46. Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics. 2010 Sep 1;26(17):2204–7.

47. Corces MR, Granja JM, Shams S, Louie BH, Seoane JA, Zhou W, et al. The chromatin accessibility landscape of primary human cancers. Science. 2018 Oct 26;362(6413):eaav1898.

48. Sollis E, Mosaku A, Abid A, Buniello A, Cerezo M, Gil L, et al. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. Nucleic Acids Res. 2023 Jan 6;51(D1):D977–85.

Figure 1. A schematic illustrating the data curation, web construction, and functional modules of scCancerExplorer.

Figure 2. An example for exploring an interested gene (*MISP*) with the "Integrated multi-omics analysis" module.

(A) The gene expression levels of *MISP* in normal tissues and cancer tissues in TCGA STAD dataset.

(B) The relative expression levels of *MISP* across various cell types of STAD with single-cell RNA-seq dataset (GSE210347).

(C) TCGA survival analysis shows that the patients with lower MISP expression levels (<median value) have significantly poorer prognosis.

(D) The motif activities of potential transcript factors for *MISP* deduced from single-cell ATAC-seq data (GSE173682).

(E) The gene activity of *MISP* across different cell types deduced from single-cell ATAC-seq data (GSE173682).

(F) The single-cell DNA methylation levels of *MISP* promoter among normal epithelial cells, cancer cells from primary tumors, and cancer cells from metastasis, which were calculated with single-cell whole-genome DNA methylation sequencing data (HRA001689).

TCGA, The Cancer Genome Atlas; STAD, stomach adenocarcinoma.

Figure 3. Examples for using the "Pan-cancer analysis" and "Query gene expression" functions within "Single-cell transcriptome" module.

(A) The dot plot shows the expression levels and expressed percentages of selected genes in cancer cells of primary tumors from multiple cancer types.

(B) The dot plot shows the expression patterns of selected gene in both the single-cell pan-cancer data in our database and the data uploaded by users.

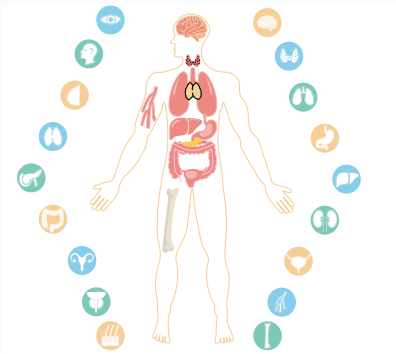
(C) Querying the relative expression levels of *EPCAM* across various cell types with "Query gene expression" function in colorectal cancer dataset HRA000201

488 (D) Other related information, such as cell type identities, sampling locations, cancer
489 stages, cell numbers and fractions of each cell type, expression levels of cell cycle
490 genes, tables for differentially expressed genes, were also provided in the "Query gene
491 expression" function.
492 UMAP, Uniform Manifold Approximation and Projection; NSCLC, non-small cell
493 lung cancer; CHOL, cholangiocarcinoma; STAD, stomach adenocarcinoma; PAAD,
494 pancreatic adenocarcinoma; KICH, kidney chromophobe.
495
496 **Figure 4. Examples for using the "DIY expression plot" function within "Single-**
497 **cell transcriptome" module.**
498 (A) With the "DIY expression plot" function, the relative expression levels of multiple
499 selected genes among multiple cell groups can be shown in multiple figure types,
500 including dot plot (top), violin plot (middle), and heatmap (bottom).
501 (B-D) By switching options in the "Group by" select box, the expression levels among
502 different sampling locations (e.g. primary tumor v.s. metastasis) (B), different cancer
503 stages (e.g. stage I v.s. stage III) (C), and various cancer subtypes (e.g. poorly-
504 differentiated adenocarcinoma v.s. well-differentiated adenocarcinoma) (D) can be
505 plotted.
506 COAD, colon adenocarcinoma; READ, rectal adenocarcinoma.
507
508 **Figure 5. Examples for using "Single-cell epigenome" module.**
509 (A) The "Tanghulu" plot displays DNA demethylation regions in colorectal cancer cells
510 (dataset ID: GSE97693), with each circle representing a CpG site. White dots represent
511 unmethylated sites, while black dots represent methylated sites. Gray dots at the bottom
512 indicate covered CpG sites within selected regions in selected cells.
513 (B) The DNA methylation levels of *CALHM3* promoter for normal epithelial cells,
514 cancer cells from primary tumors, and cancer cells from metastasis from gastric cancer
515 patients (dataset ID: HRA001689).
516 (C) Screenshot of the epigenome browser showing the chromatin accessibility levels

517 around the transcription start site of *LAPTM4B* across multiple cell types (dataset ID:
518 GSE173682). The gray regions illustrate the potential enhancers regulating *LAPTM4B*
519 in cancer cells.
520 (D) Querying the motif activity of transcription factor LEF1 across various cell types
521 in kidney renal clear cell carcinoma (dataset ID: GSE181062) with the "TF motif
522 activity" function.
523 (E) Querying the motif activity of multiple transcription factor (TBX5, RREB1, and
524 LEF1) in multiple cell types (dysfunctional CD8+ T cells, CD8+ effector T cells, and
525 naive CD4+ T cells) in kidney renal clear cell carcinoma (dataset ID: GSE181062) with
526 "DIY ATAC-seq plot" function.
527
528 **Figure 6. Examples for using "Single-cell genome" module.**
529 (A) and (B). The heatmaps showing the copy number variations of STAD patients
530 (dataset HRA001689) (A) and COAD & READ patients (dataset GSE97693) (B).
531 STAD, stomach adenocarcinoma; COAD, colon adenocarcinoma; READ, rectal
532 adenocarcinoma.
533
534 **Figure 7. Examples for using the "Survival analysis" and "ATAC-seq browser"**
535 **functions within "TCGA analysis" module.**
536 (A) Survival analysis of all COAD patients according to *TPM2* expression levels.
537 (B) and (C). Only the COAD patients with *APC* mutation (B) or diagnosed as
538 adenocarcinoma (C) were selected to perform survival analysis according to *TPM2*
539 expression levels.
540 (D) and (E). Browsing the chromatin accessibility pattern of *MISP* (D) *and SRI* (E) in
541 TCGA STAD ATAT-seq data with the "ATAC-seq browser" function.
542 TCGA, The Cancer Genome Atlas; ATAC, assay for transposase-accessible chromatin;
543 COAD, colon adenocarcinoma; STAD, stomach adenocarcinoma.

Figure 1

Data curation



>4 million
Single cells

51
Cancer types

4
Omics

5
Modules

12
Functions

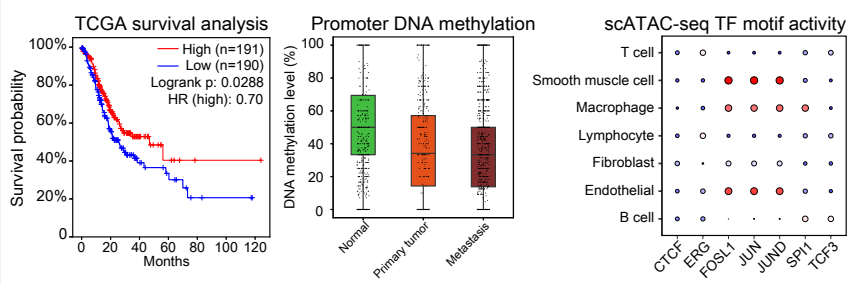
Web construction

- **Database** — PostgreSQL
- **Server** — Python, django
- **Front** — Angular, NG-ZORRO, Nginx

Functional modules

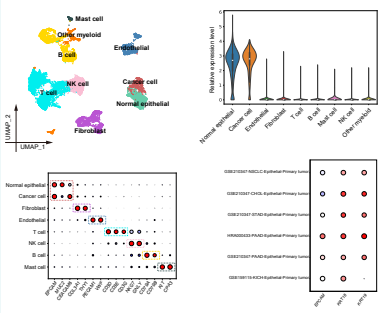
1 Integrated multi-omics analysis

- TCGA RNA-seq
- Single-cell RNA-seq
- TCGA survival analysis
- TF motif activity (scATAC-seq)
- Single-cell DNA methylation



2 Single-cell transcriptome

- Query gene expression
- DIY expression plot
- Pan-cancer analysis

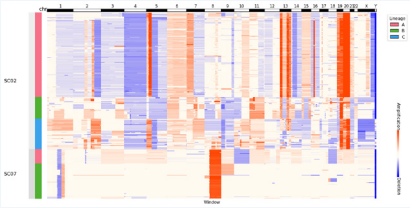


3 Single-cell epigenome

- Focal DNAme plot
- Promoter DNAme plot
- Epigenome browser
- TF motif activity
- DIY ATAC-seq plot

4 Single-cell epigenome

- Copy number variation



5 TCGA + GWAS

- Survival analysis (mutation + gene expression)
- ATAC-seq peak
- Risk SNP

Figure 2

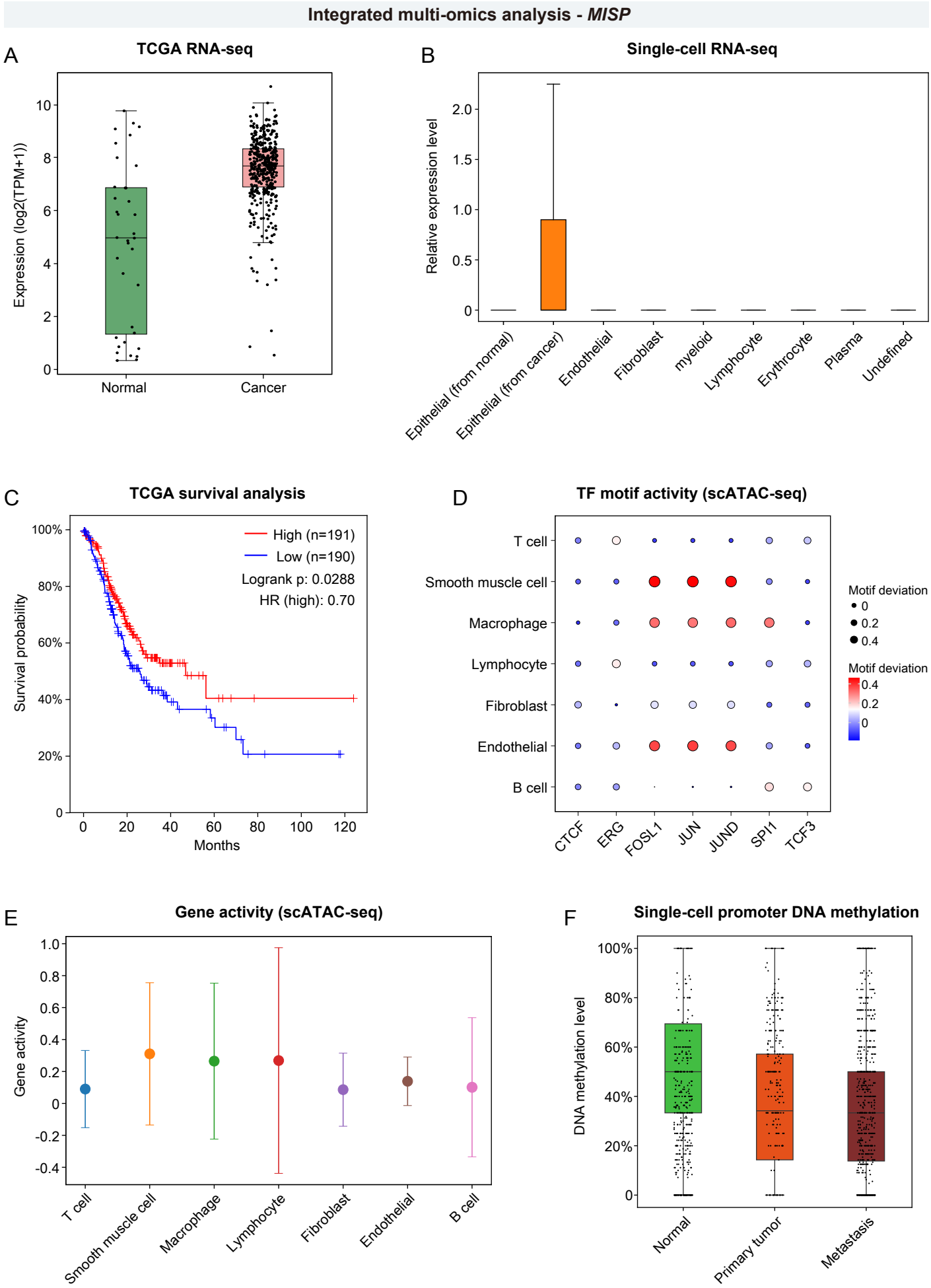

Integrated multi-omics analysis - MISP
A
TCGA RNA-seq
Expression (log2(TPM+1))
Normal
Cancer
B
Single-cell RNA-seq
Relative expression level
Epithelial (from normal)
Epithelial (from cancer)
Endothelial
Fibroblast
myeloid
Lymphocyte
Erythrocyte
Plasma
Undefined
C
TCGA survival analysis
High (n=191)
Low (n=190)
Logrank p: 0.0288
HR (high): 0.70
Survival probability
Months
D
TF motif activity (scATAC-seq)
T cell
Smooth muscle cell
Macrophage
Lymphocyte
Fibroblast
Endothelial
B cell
CTCF
ERG
FOSL1
JUN
JUND
SPI1
TCF3
Motif deviation
Motif deviation
E
Gene activity (scATAC-seq)
Gene activity
T cell
Smooth muscle cell
Macrophage
Lymphocyte
Fibroblast
Endothelial
B cell
F
Single-cell promoter DNA methylation
DNA methylation level
Normal
Primary tumor
Metastasis

Figure 3

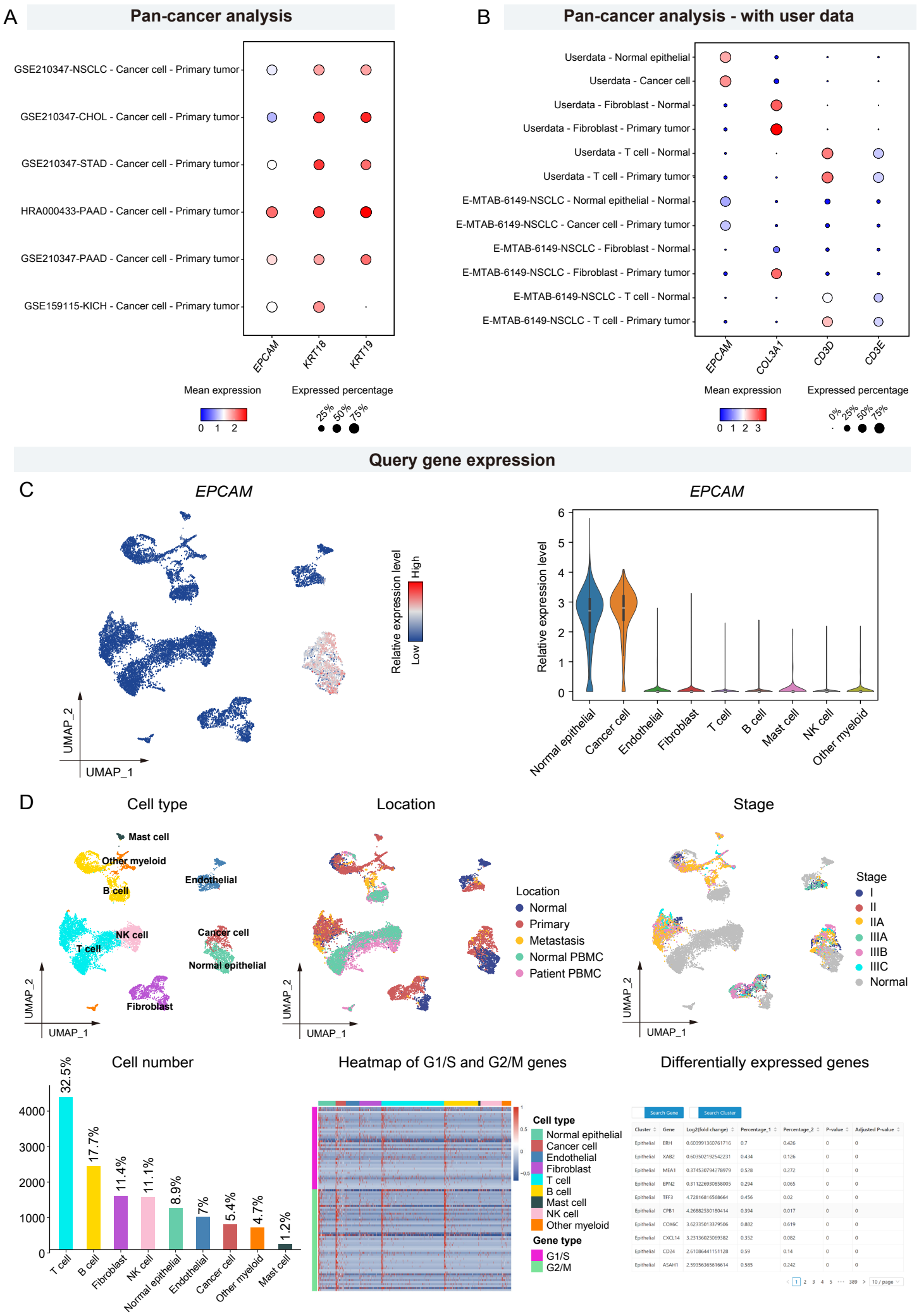

A
Pan-cancer analysis
GSE210347-NSCLC - Cancer cell - Primary tumor
GSE210347-CHOL - Cancer cell - Primary tumor
GSE210347-STAD - Cancer cell - Primary tumor
HRA000433-PAAD - Cancer cell - Primary tumor
GSE210347-PAAD - Cancer cell - Primary tumor
GSE159115-KICH - Cancer cell - Primary tumor
EPCAM
KRT18
KRT19
Mean expression
Expressed percentage
0 1 2
25% 50% 75%
B
Pan-cancer analysis - with user data
Userdata - Normal epithelial
Userdata - Cancer cell
Userdata - Fibroblast - Normal
Userdata - Fibroblast - Primary tumor
Userdata - T cell - Normal
Userdata - T cell - Primary tumor
E-MTAB-6149-NSCLC - Normal epithelial - Normal
E-MTAB-6149-NSCLC - Cancer cell - Primary tumor
E-MTAB-6149-NSCLC - Fibroblast - Normal
E-MTAB-6149-NSCLC - Fibroblast - Primary tumor
E-MTAB-6149-NSCLC - T cell - Normal
E-MTAB-6149-NSCLC - T cell - Primary tumor
EPCAM
COL3A1
CD3D
CD3E
Mean expression
Expressed percentage
0 1 2 3
0% 25% 50% 75%
Query gene expression
C
EPCAM
UMAP_2
UMAP_1
Relative expression level
High
Low
EPCAM
Relative expression level
Normal epithelial
Cancer cell
Endothelial
Fibroblast
T cell
B cell
Mast cell
NK cell
Other myeloid
D
Cell type
Mast cell
Other myeloid
B cell
Endothelial
NK cell
T cell
Cancer cell
Normal epithelial
Fibroblast
UMAP_2
UMAP_1
Location
Location
Normal
Primary
Metastasis
Normal PBMC
Patient PBMC
Stage
Stage
I
II
IIA
IIIA
IIIB
IIIC
Normal
Cell number
32.5%
17.7%
11.4%
11.1%
8.9%
7%
5.4%
4.7%
1.2%
T cell
B cell
Fibroblast
NK cell
Normal epithelial
Endothelial
Cancer cell
Other myeloid
Mast cell
Heatmap of G1/S and G2/M genes
Cell type
Normal epithelial
Cancer cell
Endothelial
Fibroblast
T cell
B cell
Mast cell
NK cell
Other myeloid
Gene type
G1/S
G2/M
Differentially expressed genes
Search Gene
Search Cluster
Cluster Gene Log2(fold change) Percentage_1 Percentage_2 P-value Adjusted P-value
Epithelial ERH 0.603991360761716 0.7 0.426 0 0
Epithelial XAB2 0.603502192542231 0.434 0.126 0 0
Epithelial MEA1 0.374530794278979 0.528 0.272 0 0
Epithelial EPN2 0.311226930858005 0.294 0.065 0 0
Epithelial TFF3 4.72816816568664 0.456 0.02 0 0
Epithelial CPB1 4.26882530180414 0.394 0.017 0 0
Epithelial COX6C 3.62335013379506 0.882 0.619 0 0
Epithelial CXCL14 3.23136025069382 0.352 0.082 0 0
Epithelial CD24 2.61086441151128 0.59 0.14 0 0
Epithelial ASAH1 2.59356365616614 0.585 0.242 0 0
1 2 3 4 5 ··· 389
10 / page

Figure 4

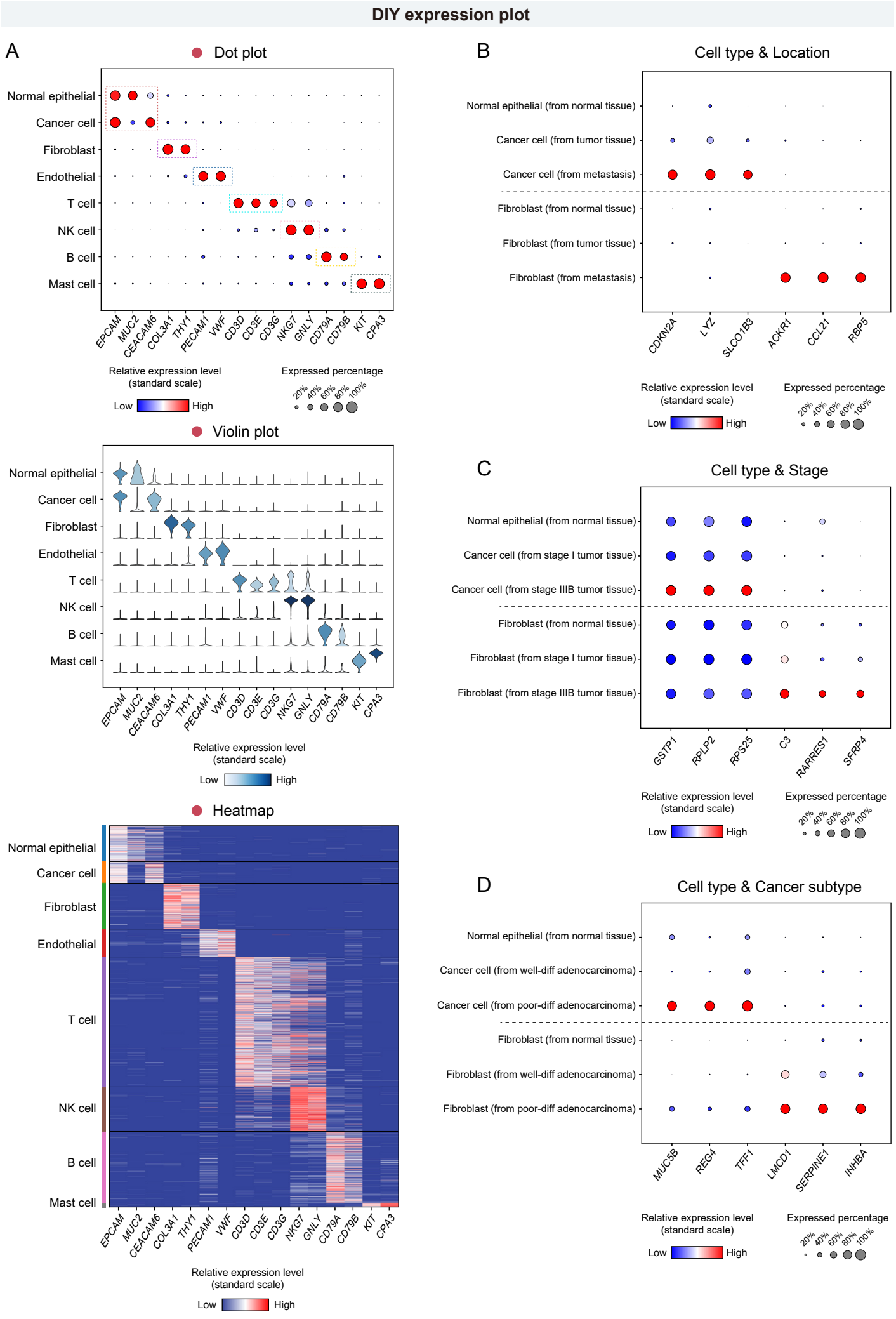

DIY expression plot
A
Dot plot
Normal epithelial
Cancer cell
Fibroblast
Endothelial
T cell
NK cell
B cell
Mast cell
EPCAM MUC2 CEACAM6 COL3A1 THY1 PECAM1 VWF CD3D CD3E CD3G NKG7 GNLY CD79A CD79B KIT CPA3
Relative expression level (standard scale)
Low High
Expressed percentage
20% 40% 60% 80% 100%
Violin plot
Heatmap
B
Cell type & Location
Normal epithelial (from normal tissue)
Cancer cell (from tumor tissue)
Cancer cell (from metastasis)
Fibroblast (from normal tissue)
Fibroblast (from tumor tissue)
Fibroblast (from metastasis)
CDKN2A LYZ SLCO1B3 ACKR1 CCL21 RBP5
C
Cell type & Stage
Cancer cell (from stage I tumor tissue)
Cancer cell (from stage IIIB tumor tissue)
Fibroblast (from stage I tumor tissue)
Fibroblast (from stage IIIB tumor tissue)
GSTP1 RPLP2 RPS25 C3 RARRES1 SFRP4
D
Cell type & Cancer subtype
Cancer cell (from well-diff adenocarcinoma)
Cancer cell (from poor-diff adenocarcinoma)
Fibroblast (from well-diff adenocarcinoma)
Fibroblast (from poor-diff adenocarcinoma)
MUC5B REG4 TFF1 LMCD1 SERPINE1 INHBA

Figure 5

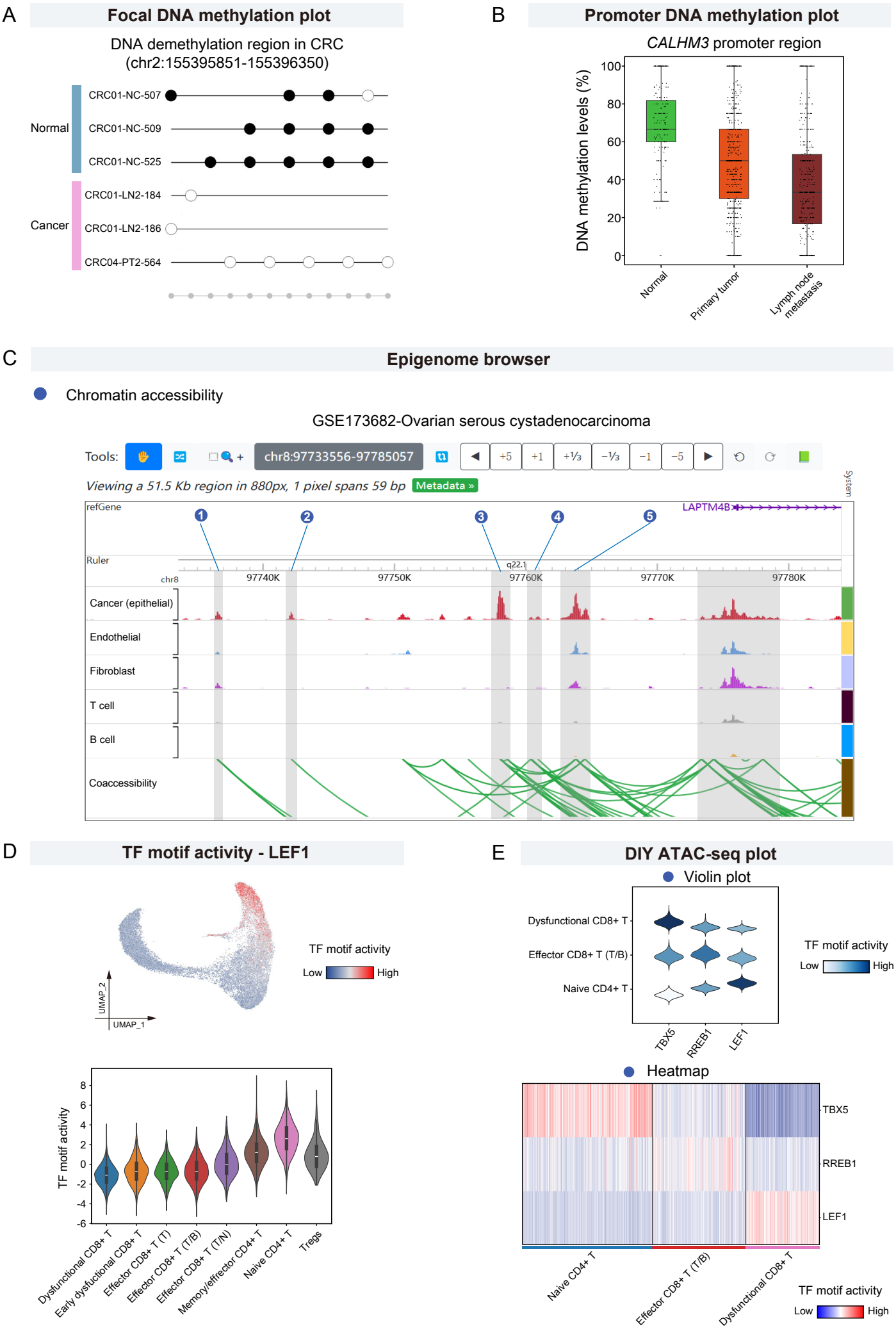

A
Focal DNA methylation plot
DNA demethylation region in CRC
(chr2:155395851-155396350)
Normal
CRC01-NC-507
CRC01-NC-509
CRC01-NC-525
Cancer
CRC01-LN2-184
CRC01-LN2-186
CRC04-PT2-564
B
Promoter DNA methylation plot
CALHM3 promoter region
DNA methylation levels (%)
100
80
60
40
20
0
Normal
Primary tumor
Lymph node metastasis
C
Epigenome browser
Chromatin accessibility
GSE173682-Ovarian serous cystadenocarcinoma
Tools:
chr8:97733556-97785057
+5
+1
+1/3
−1/3
−1
−5
Viewing a 51.5 Kb region in 880px, 1 pixel spans 59 bp
Metadata »
System
refGene
LAPTM4B
Ruler
q22.1
chr8
97740K
97750K
97760K
97770K
97780K
Cancer (epithelial)
Endothelial
Fibroblast
T cell
B cell
Coaccessibility
D
TF motif activity - LEF1
UMAP_2
UMAP_1
TF motif activity
Low
High
TF motif activity
8
6
4
2
0
-2
-4
-6
Dysfunctional CD8+ T
Early dysfuctional CD8+ T
Effector CD8+ T (T)
Effector CD8+ T (T/B)
Effector CD8+ T (T/N)
Memory/effector CD4+ T
Naive CD4+ T
Tregs
E
DIY ATAC-seq plot
Violin plot
Dysfunctional CD8+ T
Effector CD8+ T (T/B)
Naive CD4+ T
TBX5
RREB1
LEF1
TF motif activity
Low
High
Heatmap
TBX5
RREB1
LEF1
Naive CD4+ T
Effector CD8+ T (T/B)
Dysfunctional CD8+ T
TF motif activity
Low
High

Figure 6

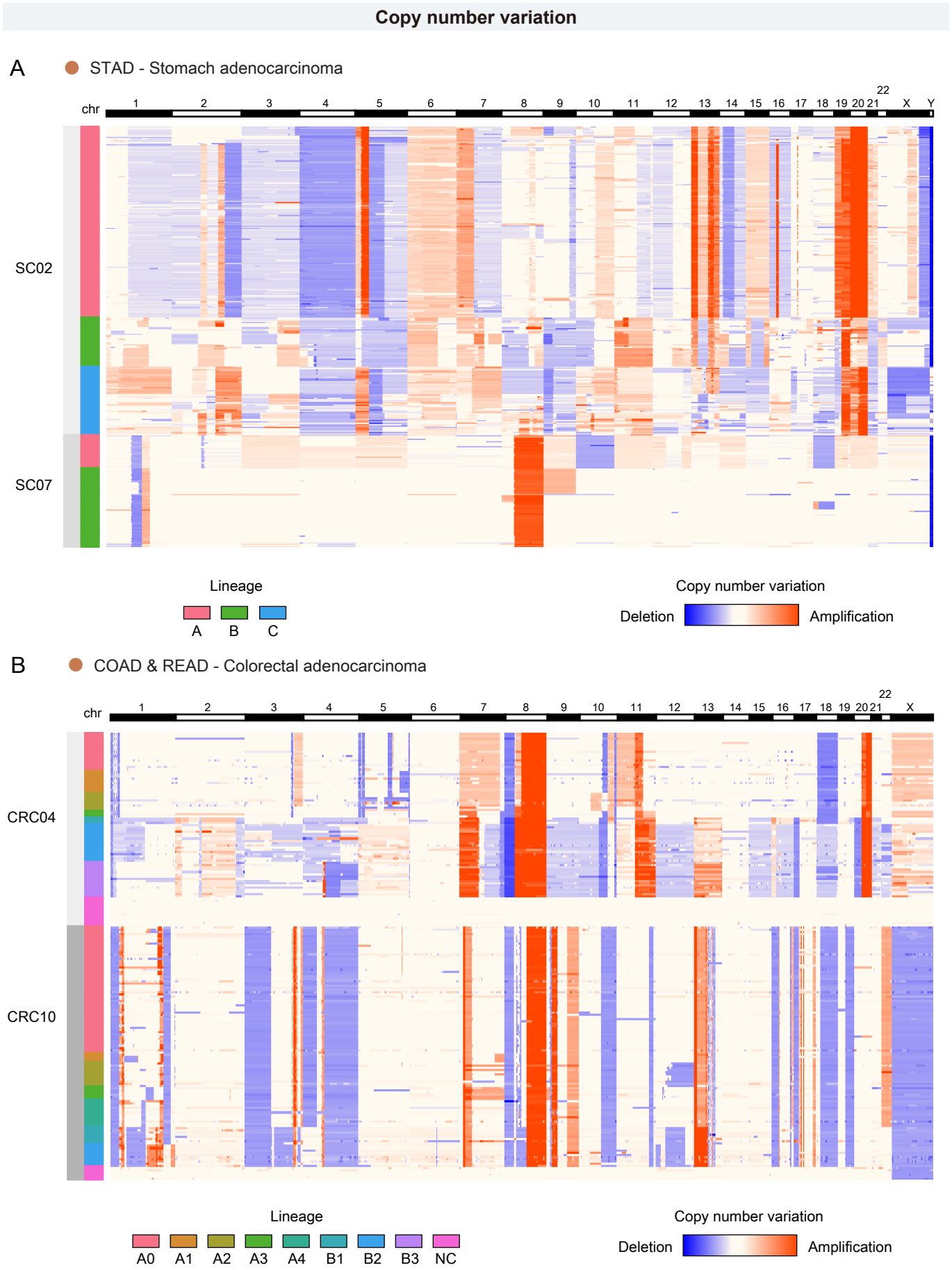
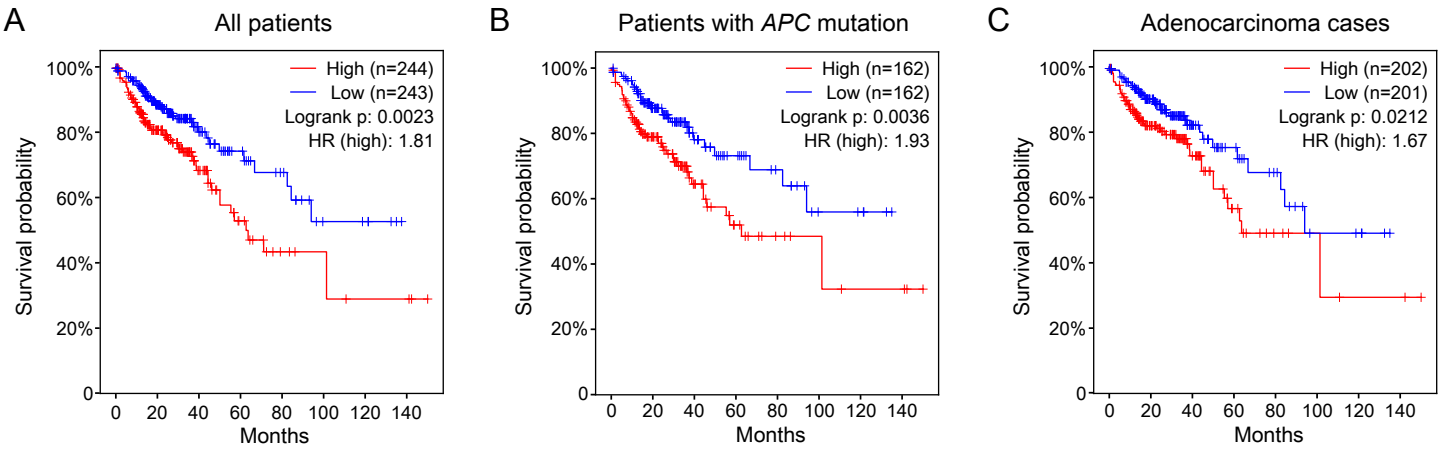

Copy number variation
A
STAD - Stomach adenocarcinoma
chr
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
SC02
SC07
Lineage
A B C
Copy number variation
Deletion
Amplification
B
COAD & READ - Colorectal adenocarcinoma
chr
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
CRC04
CRC10
Lineage
A0 A1 A2 A3 A4 B1 B2 B3 NC
Copy number variation
Deletion
Amplification

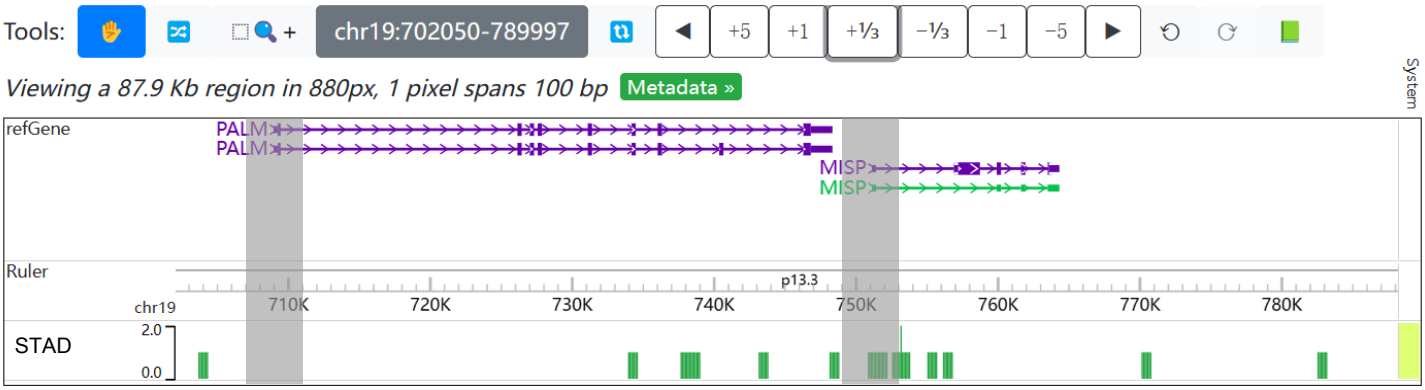
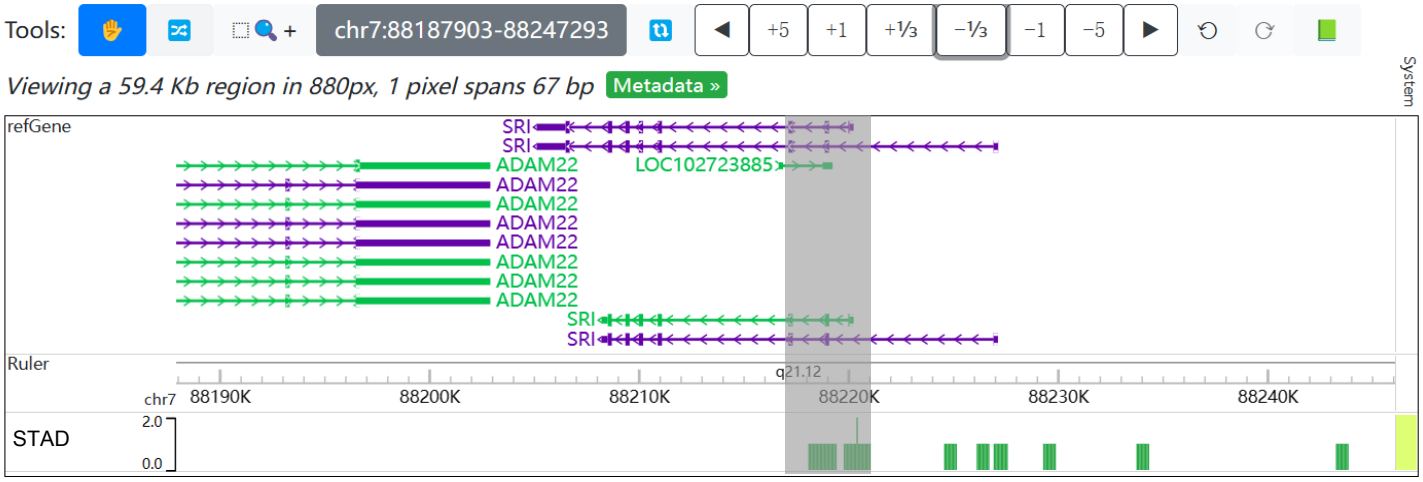
Figure 7

Survival analysis - *TPM2*



ATAC-seq browser

D Chromatin accessibility - *MISP*



E Chromatin accessibility - *SRI*